Cluster Analysis of Real Time Location Data - An Application of Gaussian Mixture Models

Alvaro Ortiz-Vazquez
EdLab
Teachers College Columbia University
New York, New York USA
ao2444@columbia.edu

Xiang Liu
EdLab
Teachers College Columbia University
New York, New York USA
xl2438@tc.columbia.edu

Ching-Fu Lan
EdLab
Teachers College Columbia University
New York, New York USA
cl2483@tc.columbia.edu

Hui Soo Chae
EdLab
Teachers College Columbia University
New York, New York USA
hsc2001@tc.columbia.edu

Gary Natriello
EdLab
Teachers College Columbia University
New York, New York USA
gjn6@tc.columbia.edu

ABSTRACT
Clustering analysis in the context of education is important for determining the effectiveness of group activities especially when participants freely rotate between groups such as in a gallery exhibit or other informal learning space or set-up. In this paper, we cover a method of applying Gaussian Mixture Models to two-dimensional data. We further describe the analysis procedure, and the success of implementing this analysis using simulated data and real data. Finally, we discuss some educational applications as well as future directions for this research.

Keywords
Gaussian Mixture Models, MCMC, Gibbs Sampling, Real-Time Location System, Informal Learning Spaces, Learning Analytics, Dynamic Mixture Model

1. INTRODUCTION

Real-time locating systems have become increasingly popular and are predicted to be more widely adopted in informal learning institutions such as libraries, museums, and after school spaces in the next few years [2] [4]. Location intelligence and contextually relevant information can inform dynamically customized information and meaningful learning analytics for both learners and educators based on visitors and/or learners' location [3]. Such data are especially useful to understand social interactions in informal learning events. Therefore, it is essential for researchers to develop data mining methods to more efficiently and effectively explore real-time location data of learners.

Gaussian Mixture Models (GMM) are very useful for analyzing two-dimensional data which may be clustered into groups such as that collected by a real-time locating system in an informal learning space. To estimate the parameters of the GMM we employ a Markov Chain Monte Carlo method of Gibbs sampling [1] whose stationary state is the posterior distribution of the mixture model. This method applied to a frozen snapshot of the two-dimensional real-time location tracking data allows us to gain information about the groups, such as group membership, group location, and internal group dispersion, based only on the tag position data. Other algorithms such as k-means clustering may similarly cluster two-dimensional data but are non-parametric whereas Gibbs sampling is parametric.

2. DATA ANALYSIS

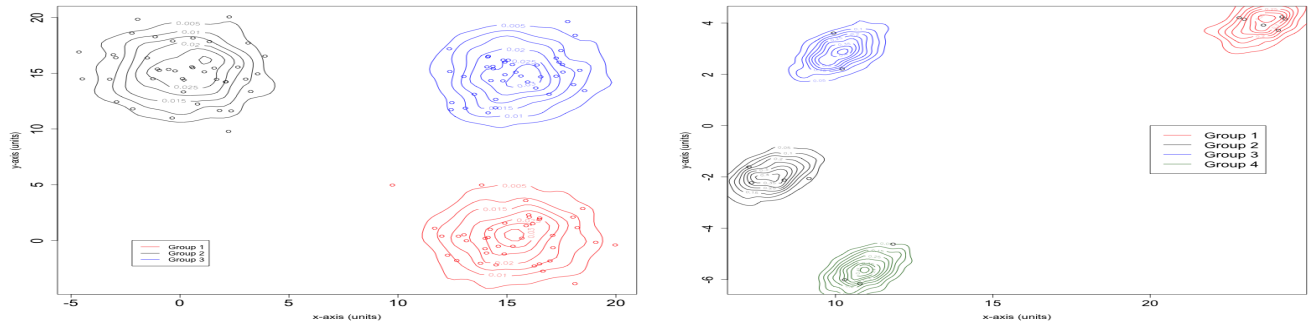
2.1 Simulations

To test the Gibbs sampling process and our R code we have drawn a set of location data points from bivariate normal distributions centered around three different centers ($\boldsymbol{\mu}_1 = (15, 15)$, $\boldsymbol{\mu}_2 = (15, 0)$,$\boldsymbol{\mu}_3 = (0, 15)$) with a common covariance. We observed the latent parameters of our Gibbs sampler reaching a stationary state in less than 100 iterations. In Figure 1a we generate estimated points using the estimated group centers and covariance and perform kernal density estimates to generate the coverage contours plotted over the original generated data. The percentage of estimated points outside the contours is marked on the contour lines. In this case we see that for 120 data points, a small number of the data lie outside of the 99.5% percent coverage contours. We can also verify the results by comparing the generating values for the centers and covariance with the estimated values.

2.2 Applications on Real Data

The real data were collected at an Edlab meeting at an innovative learning space: the Smith Learning Theater at Teachers College Columbia University. The Smith Learning Theater features technologies such as the Quuppa ^{TM} real-time locating system, installed to return measurable results and provide feedback to organizers and facilitators. In this meeting, 15 EdLab members wore Quuppa real-time locating tags and freely explored four stations of augmented/virtual reality apps in order to provide reviews for a national edtech competition. Applying the Gibbs Sampling method over the real data we again obvserved convergence within just 100 iterations. Again the coverage contours are drawn onto the plot of the positions in Figure 1b. In this analysis we did not have previous knowledge on the station device locations likely to be correlated with the group centers. However, we can still verify the success of the algorithm by noting that the data points are largely within the ninety-five percent coverage region. As such, our method returns accurate group information even with a small dataset.

3. DISCUSSION



(a) (b)

Figure 1: Kernal Density Estimate Contour Plots Over Simulation Data (a) and Real Data (b)

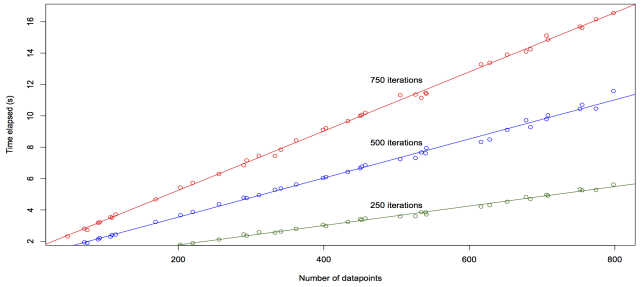


Figure 2: Linear Correlation Between Computational Time and the Number of Data Points

3.1 Educational Research and Applications

Our method has the limitation that the expected number of groups must be specified prior to performing the Gibbs sampling. This quantity can be available for events where group work takes place, or participants move around through different stations. In such an event our analysis can be implemented repeatedly over a series of consecutive discrete snapshots covering a period of time. By observing the group membership at each snapshot, the educator can determine information about who moved together as a group, or who moved mostly independently. Common group membership can be denoted in an adjacency matrix for the tags where the value for each index (i, j) is the number of snapshots in which two locating tags y_i, y_j shared the same group assignment. This approach has the potential to provide information about whether the learning space or activity was better suited for group learning or independent learning and the preferences of each participant to remain with the same group of people or move about with different people. In other events where group work may be taking place one can easily determine the amount of cross-group collaboration during a period of time by again looking at the cumulative group assignment data.

3.1.1 *Feasibility Analysis*

The implementation of the Gibbs Sampling algorithm takes linear $\mathcal{O}(N)$ time where N is the number of position data points in a single snapshot. We can generate N position data points and record the time elapsed for M iterations and visualize the linear relationship in Figure 2. Given an hour long event with 500 participants, covered by 360 snapshots, the linear model suggests that one could perform 250 iterations of the sampler over every snapshot in under twenty minutes. As such implementation of our method is feasable for most educational contexts.

3.2 Future Work

While our model is useful to see the group information within a snapshot of real-time location data, we believe that more important data will arise from extending our current mixture model to a Dynamic Mixture Model (DMM) [5]. In such a DMM, the group distribution of each snapshot would be dependent on the previous one. According to Wei et al. (2007) the assumption that two consecutive snapshots are dependent can allow us to analyze important patterns that would otherwise be missed in discrete snapshot analysis. By incorporating the temporal component, we expect to more accurately model transitions between groups. The application of our method is especially valuable in informal learning spaces as many learning events in these spaces encourage free exploration and group interactions, and evaluating learners' engagement and social group dynamics is challenging using other traditional research methods.

References

[1] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE transactions on pattern analysis and machine intelligence*, 6(6):721–41, jun 1984.

[2] B. Herr-Stephenson, D. Rhoten, D. Perkel, C. Sims, A. Balsamo, M. Klosterman, and S. S. Bautista. *Digital Media and Technology in Afterschool Programs , Libraries , and Museums*. 2011.

[3] K. Jaebker and G. Bowman. Context is king: Using indoor-location technology for new visitor experiences | MW2015: Museums and the Web 2015, 2017.

[4] L. Johnson, S. Adams Becker, M. Cummins, V. Estrada, A. Freeman, and C. Hall. *Horizon Report: 2016 Higher Education Edition*. The New Media Consortium, Austin, Texas, museum edi edition, 2016.

[5] X. Wei, J. Sun, and X. Wang. Dynamic mixture models for multiple time series. *Proceedings of the 20th international joint conference on Artifical intelligence*, (Dmm):2909–2914, 2007.